OPEN UNIVERSITY DATA

Martin Mitrevski, Milos Jovanovik, Riste Stojanov, Dimitar Trajanov
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
Skopje, Republic of Macedonia

ABSTRACT

Today, there is a growing trend for publishing public data in an open format, on the web, making it available for everyone to use and reuse. This idea has been widely supported by governments and companies throughout the world, which have made their own public data available in such way. Some of them, like the World Bank, even challenge developers to write applications based on their open data, by organizing competitions [1]. Data has become the new raw material of the 21st century [2]. The Linked Open Data project has begun turning the document-oriented web into a database of global proportions [3].

The Faculty of Computer Science and Engineering joins this trend by making parts of its public data available as open data. This paper introduces a system for mapping relational data from databases into data represented in semantic web format (N3 and RDF), as well as editing and querying the data by using a SPARQL endpoint. Here we describe the process of publishing the open data from our Faculty, as well as some basic information from the other faculties which are part of the "Ss. Cyril and Methodius" University in Skopje. We also propose some possible applications which can use this open data, and in that way add more value to it.

I. INTRODUCTION

Most information systems today store their data in relational databases. The data they store usually has parts of it which are not public by their nature, so they must not be made openly available. Therefore, there is a need for a mechanism that would automatically publish data on the web in a semantic web format, by extracting only specific information from the relational databases. One problem is that the tools that are developed for mapping a database into a desired format do not provide a way to select only specific parts from the database; they usually map the entire database. Another problem is that they do not provide options to link that data to other well-known ontologies and semantically annotated data on the web. This must be done manually. This implies creating another layer on top of these tools to provide the necessary requirements. This paper discusses one possible way to do this.

Another challenge is the ontology engineering. There are a lot of available ontologies on the web, especially for describing academic environments. The ontology repository should contain ontologies that best match the scope of our data, but still link to most of the data from other universities. A balance should be made between these two choices when choosing the ontologies. Another issue is that new ontologies are created every day and it is not possible to predict which university ontology would be the most popular in the future. There should be an easy way to switch the link to other ontologies, if the chosen ones lose their wide support, because it is more useful to have your data linked with as much as possible other external entities. The flexibility of editing the ontology repository is important for another reason - what if we make changes in the database and some tables are not needed anymore, and others are added? And what if we decide to publish data that we previously considered private? A possible solution is presented in the paper.

II. RELATED WORK

As we mentioned before, the trend of publishing data on the web in an open and semantic web format is very popular in the academic environments. There are few universities, most of them in the UK, which have already started open data projects. They publish schedules, modules, subjects, phonebooks of the employees, etc. Besides them, other institutions such as the Governments in UK, USA and Netherlands, World Bank, NASA, BBC, ACM and Amazon have also made available some of their data on the web [4]. The governments publish lots of public information about their work, such as contracts, how they spend the citizens' money, who their ministers are meeting, etc [5]. There are already some excellent mobile applications in the UK using this data [5]. The data is available either in RDF, N3, Excel, XML format or as JSON objects [6].

There are many tools for achieving this, such as D2R Server [7], Oracle Spatial 11g [8], Asio Semantic Bridge [9], SquirrelRDF [10] and many others. We use the D2R Server, which enables RDF and HTML browsers to navigate the content of the database, and allows applications to query the database using the SPARQL query language [7]. The other tools will not be discussed in this paper.

III. SYSTEM ARCHITECTURE

The system for creating open and linked data from relational databases consists of five parts: the relational database, the D2R Server, the Mapping Tool, the Ontology Repository and the RDF Documents, and the SPARQL endpoint, called Snorql, for previewing the data (Figure 1).

In a nutshell, the Open Data Database gets data from the FCSE DB. The data is mapped by the D2R Engine, generating an N3 file, which is then converted by the Mapping Tool into an RDF file. The RDF file can be expanded to xml-like tree. Then the tree can be annotated with the ontologies from the Repository very easily, just by right clicking on the element we want to annotate and selecting one of the ontologies we added to the repository. Then the RDF file is stored within the RDF documents Repository, and can be used later. It is also converted back

into N3 file so the D2R Server can invoke the SPARQL endpoint and show the linked structured data. The data can be queried with SPARQL Query Language and is accessible in few formats, so it can also be extracted and used in other applications.

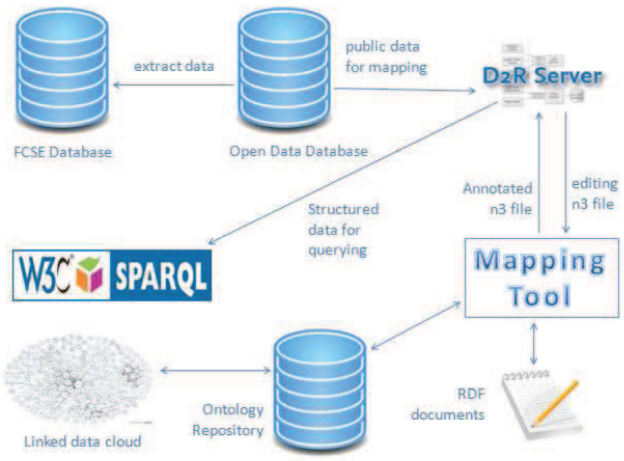

Figure 1: System Architecture.

We will go into more details on the parts of the system further in this paper.

IV. Database

The data of the Faculty of Computer Science and Engineering is stored in a relational database on a Microsoft SQL Server. It contains confidential information about the employees and the students, which must not be published on the web. To protect this sensitive data, the D2R server should not do the mapping directly on this database, because it is not secure and reliable. For this reason, the Open Data DB is created. This database contains only the data which we want to make public. It extracts data from the original database, and is updated whenever there are changes in the source database, using SQL Queries. In this way, we manage to avoid consistency problems.

This database contains information about all the faculties in our university such as names, addresses, phones, email and web pages. We publish more information about our Faculty - Institutes, Modules, Programs, Courses, Subjects and Employees.

The D2R Server just maps the tables from this database and the generated N3 file contains only the mapping, and not the data. The data is retrieved from the Open Data DB, which means that changes in the database will reflect on the data in the SPARQL endpoint. There is no need to store the Semantic Data in a dedicated RDF triple store, because requests from the Web are rewritten into SQL queries via the mapping. This on-the-fly translation allows publishing of RDF from large live databases [7].

V. The Mapping Tool

The D2R Server lacks an interface which can enable us to link ontologies to the tables. Therefore, the data that is previewed in the SPARQL Endpoint is just open structured data, it is not linked to the Linked Data Cloud and is not very useful, because Semantic Web Browsers cannot associate it with already existing data types to provide context. The D2RQ Platform only enables manually editing the generated N3 file using the D2RQ Mapping Language [7].

Manually editing the N3 file can be time consuming, prone to errors, ineffective and impractical, especially when there are a lot of tables to annotate. These reasons led us to create the Mapping Tool, which is a web application that utilizes the D2R Server functionalities to provide an easier and simpler way to connect the data with existing ontologies (Figure 2). It is a modification of an application for annotating web services [11]. The application can support the whole process of entering a database, generating an N3 file via the D2R Server, converting the N3 to RDF file, and then after the user annotates the file, converting it back to N3 and previewing it in the endpoint. The Mapping Tool can also accept already generated N3 or RDF files for annotation. Ontologies can be easily added to the Ontology Repository. The user can choose some class or property from the visual tree and add or remove references.

The conversion from N3 to RDF file and vice versa is done using the conversion functions of the SemWeb Library [12]. XSL transformations are used for parsing tags of the ontology into objects which are needed for the dynamic creation of the interface for selecting ontology. The D2R Server is started from the application as an external process with its batch file. The xml-like tree is dynamically generated using the classes from the Microsoft .Net library for working with XML. Adding reference to an element of the tree is done with the Stream Writer for data transfer into byte streams. The substitution of the old element with the new annotated element is done using the String Builder .NET class.

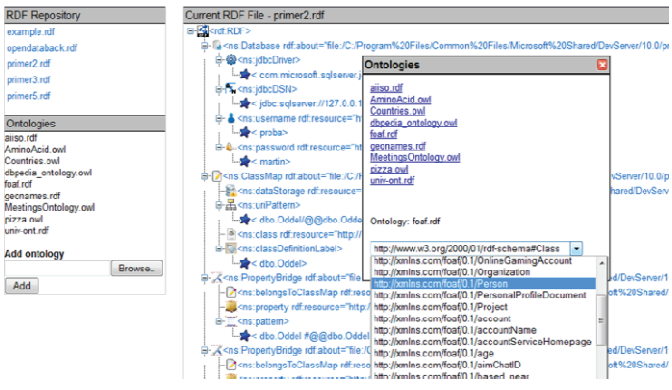

Figure 2: The Mapping Tool.

In the left upper corner is the RDF Repository, which contains the RDF files which are expanded in the RDF tree in the middle. Below the RDF Repository is the Ontologies Repository, which can be expanded with new ontologies. By

right clicking on the desired RDF class or property tag on the tree, there is an interface for adding or removing annotations. If the user chooses to add an ontology, a new dialog box is shown with the ontologies extracted from the Ontologies Repository as link buttons and the user selects one of them. If there is already an annotation for that tag, it is replaced with the latest one. Above the RDF repository (not displayed on Figure 2), there is a form which enables creating mapping from scratch. The user enters the database name, username and password, the name of the generated output file and then the application finds the database from the MS SQL Server, generates both N3 and RDF files using the D2R Server, and adds the RDF file into the RDF repository so it can be annotated.

VI. Ontology Repository

We rarely have to start from scratch when defining an ontology; there is almost always an ontology available from a third party that provides at least a useful starting point for our own ontology [13]. And of course, the idea of the semantic web is to link your data with already existing data. As we mentioned before, there are universities around the world that have already started to develop tools and ontologies for opening up and linking their data. For that reason, we try as much as possible to use and combine those well known ontologies according to our needs.

The standard ontology for describing persons, their activities and their relations to other people and objects, Friend of a Friend (FOAF) [14], is used for the employees of the Faculty. For describing the internal organizational structure of the academic environment, the Academic Institution Internal Structure Ontology [15] is used. However, it has certain limitations in describing some of the properties of our data. That is why we use another similar ontology - University ontology [16], which contains those additional features required. We connect the locations of our faculties and institutes to the Semantic Cloud with the GeoNames Ontology [17] and the time features with the Timeline ontology [18]. For other relationships between the classes, the Dublin Core Metadata Vocabulary [19] is used.

These ontologies covered our database tables as shown in Figure 3. Most of the classes of the ontologies describing academic entities shared the same names as Open Data database tables, except for the Course, whose definition corresponds to the Subjects table of our database, and the Courses for that subject correspond to the definition of the CourseGroup from the ontology.

This was not the only way to create the ontology. There are some other works for this domain, such as DBpedia's ontology about universities [20]. Universities in the UK create their own ontologies to describe some of their unique characteristics. It is very easy to switch to DBpedia or these other ontologies - the annotation of the class or object property should be changed with the Mapping Tool or by manually editing the N3 file.

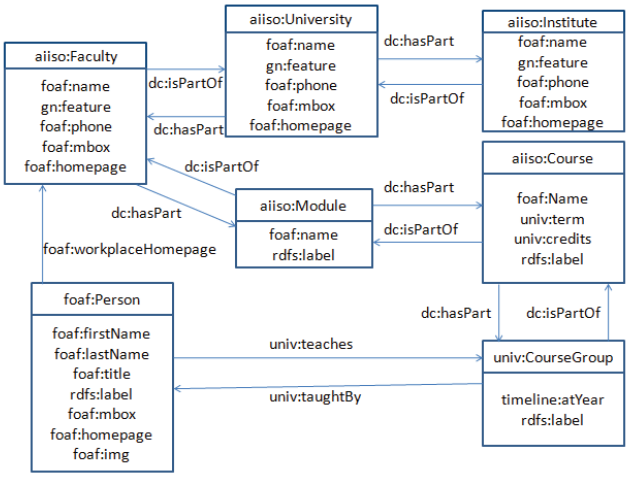

Figure 3: The database tables and properties annotated with existing ontologies. Only the data that is annotated is displayed, other properties like IDs which are mapped with the local vocabulary are not displayed.

VII. Previewing and Collecting Semantic Data

The SPARQL endpoint of the D2RQ platform [7] shows the data in triples: a subject, a predicate and an object. In the endpoint, there is a text area for writing SPARQL queries for browsing the data. The results of the queries can also be shown in JSON, XML and XML+XSL format [7], so this data can be used by both web and mobile applications. A nice feature of the SPARQL endpoint is that SPARQL queries are automatically generated while browsing the data through clicking the triples.

The data can be accessed in three ways: RDF browsers, traditional HTML browsers, and SPARQL query clients [7]. Since RDF browsers are not used widely, accessing the data with traditional browsers is a very important characteristic of the D2R Server. This is done by having XHTML representation of every resource, which is retrieved by the browsers [7]. Beside previewing and browsing, the data can be indexed using search engines.

The example in Figure 4 displays information about the subject Network Programming. It can be noted that information like credits, term, name are linked with other ontologies and information like the ID of the subject and the IDs of the courses for that subject are mapped with the local vocabulary generated by the D2R Server. From the Results dropdown list the user can choose the format of the output (triples, JSON, XML or XML+XSL).

VIII. Possible Applications

The goal of publishing our data in an open and semantically annotated format is to encourage programmers to use this data to write applications which can be very useful for our students. For example, applications with semantic search can be made to find out in which subjects the technology or research field of interest of the student is

thought, and at what level. To do this, all the subjects should be linked with the corresponding technologies. That link can provide the student with whole lot of additional resources that are connected to the cloud for that particular technology.

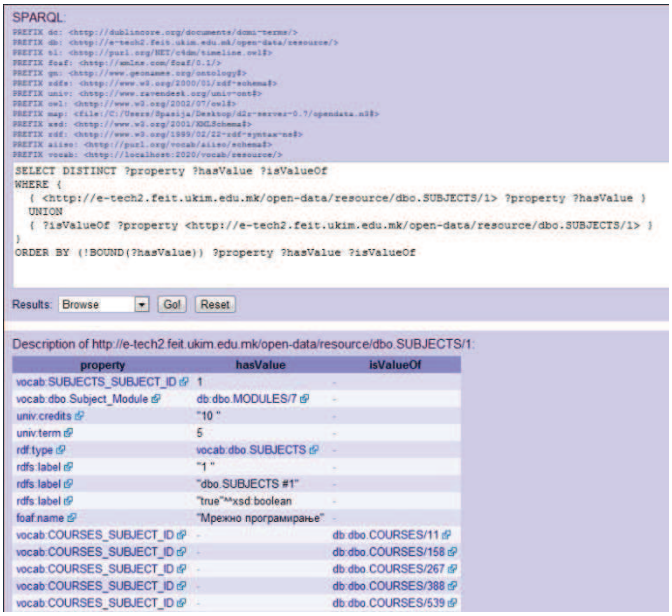

Figure 4: An example SPARQL query.

Also, if the grades of the students are made public, an intelligent portal for rating the students by their knowledge of technologies and computer skills can be made. The rating could be computed by the grades from the subjects that are connected with the skill or technology and the projects made by the student. This will enable companies to find the most suitable student to work for them.

This would also give advantage to our best students when they look for a job and would motivate the others to do a better job.

IX. Conclusion and Future Work

In this paper, we presented a mechanism for publishing data from relational databases on the Semantic Web. We introduced our Mapping Tool for annotating the data, and we described our Ontology Repository. There is still a lot of work to be done with this system. First, improvements can be made in our Ontology Repository, which should be expanded with more ontologies for the data which is still not annotated and has only meaning described with local vocabulary. Also, there is still data that is public in its nature, but is not published with this work. The Mapping Tool can be improved by using machine learning algorithms to automatically propose ontology annotations for the database tables depending on the name and data of the tables.

By opening its data, the Faculty could provide better understanding of its structure and its operations. Giving access to data can be beneficial to both the Faculty and the students. Thinking beyond the Faculty's scope, sharing data in a structured format - whether this is personal, government, scientific, medical or enterprise data has huge potential to provide solutions to all kinds of problems, so we hope this initial work will encourage other institutions in the country to follow us in publishing their data on the web in an open format.

References

[1] World Bank - Apps for Development Challenge: http://appsfordevelopment.challengepost.com/ (retrieved on 06.03.2012)

[2] Berners-Lee, T. and Shadbolt, N., *There's gold to be mined from all our data*. The Times, London, 2011.

[3] Wood, D., *Linking Enterprise Data*, Springer, Virginia, USA, 2010.

[4] Kundra, V., *Digital Fuel of the 21st Century: Innovation through Open Data and the Network Effect.* Joan Shorenstein Center on the Press, Politics and Public Policy, 2012.

[5] UK government - Opening up government: http://data.gov.uk/ (retrieved on 06.03.2012)

[6] World Wide Web Consortium (W3C) - RDF and SQL: http://www.w3.org/wiki/RdfAndSql (retrieved on 06.03.2012)

[7] Bizer, C, and Cyganiak, R., *Publishing Databases on the Semantic Web, D2R Server Technical Note,* University of Berlin, 2007.

[8] Oracle Spatial 11g: http://www.oracle.com/technetwork/database/options/spatial/overview/introduction/index.html

[9] Asio Semantic Bridge for Relational Databases: http://bbn.com/technology/knowledge/asio_sbrd

[10] SquirrelRDF: http://jena.sourceforge.net/SquirrelRDF/

[11] K. Budinoski, M. Jovanovik, R. Stojanov, D. Trajanov , *An Application For Semantic Annotation Of Web Services*, 7th International Conference for Informatics and Information Technology, February 2010.

[12] SemWeb.NET: Semantic Web/RDF Library for C#/.NET: http://razor.occams.info/code/semweb/ (retrieved on 06.03.2012)

[13] Antoniou, G., and Van Harmelen, F., *A Semantic Web primer*. The MIT Press Cambridge, Massachusetts, London, 2008, page 226.

[14] Brickley, D., and Miller, D., *FOAF Vocabulary Specification 0.98*, 2010.

[15] Academic Institution Internal Structure Ontology (AIISO): http://vocab.org/aiiso/schema (retrieved on 06.03.2012)

[16] Moving towards a university ontology: http://74-220-208-99.bluehost.com/blog/moving-toward-a-university-ontology (retrieved on 06.03.2012)

[17] GeoNames ontology: http://www.geonames.org/ontology/documentation.html (retrieved on 06.03.2012)

[18] The timeline ontology:
http://motools.sourceforge.net/timeline/timeline.html (retrieved on 06.03.2012)

[19] Dublin Core Metadata Initiative:
http://dublincore.org/documents/dcmi-terms/ (retrieved on 06.03.2012)

[20] The DBpedia Data Set: http://wiki.dbpedia.org/Datasets
(retrieved on 06.03.2012)